

TRECVID 2017 Workshop

National Institute of Standards and Technology

Multimedia Event Detection (MED) Task

Nov. 13, 2017

David Joy, Jonathan Fiscus, Andrew Delgado, Willie McClinton*

* Summer Undergraduate Research Fellow (SURF)

MED Session Schedule

11:40 – 2:40	Monday, Nov. 13
11:40 – 12:00	MED Task Overview
12:00 – 1:40	<i>Lunch</i>
1:40 – 2:00	TokyoTech+AIST (Tokyo Institute of Technology, National Institute of Advanced Industrial Science and Technology)
2:00 – 2:20	MediaMill (University of Amsterdam)
2:20 – 2:40	MED Discussion
2:40 – 3:00	<i>Break</i>

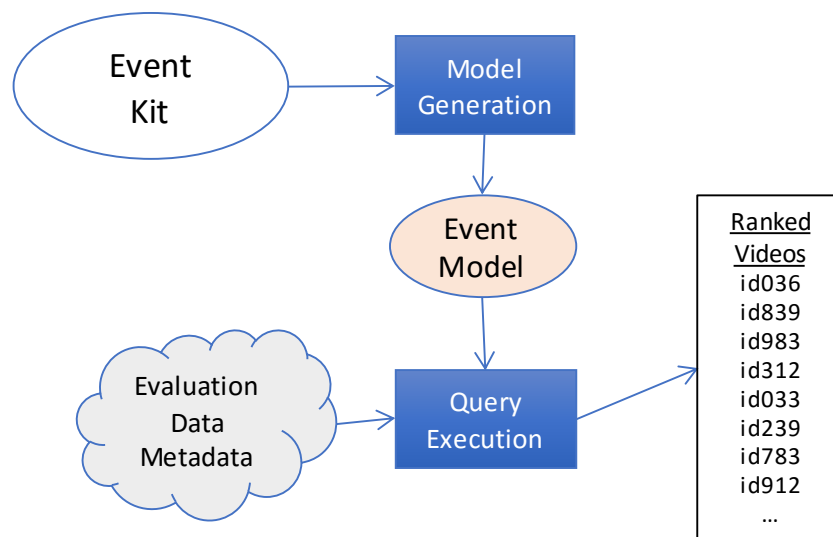
Multimedia Event Detection Task

Multimedia Event Detection (MED)

Quickly find instances of events in a large collection of search videos

A MED event is a complex activity occurring at a specific place and time involving people interacting with other people and/or objects

Notional System Diagram



Evaluation Conditions

Execution Hardware Reporting

3 Classes of Computing Hardware

- Small
 - 100 Central Processing Unit (CPU) cores, 1,000 Graphics Processing Unit (GPU) cores
- Medium
 - 1,000 CPU cores, 10,000 GPU cores
- Large
 - 3,000 CPU cores, 30,000 GPU cores

Query Training Conditions

- Pre-Specified (PS)
 - 10 Events; 10 Exemplars each
- Ad-Hoc (AH)
 - 10 Events; 10 Exemplars each

MED '17 Overview

- MED evaluations from 2010 through 2015
 - Supported by the Intelligence Advanced Research Projects Activity (IARPA) Aladdin Program and Linguistic Data Consortium (LDC) collected data
- MED 2016
 - Introduced a 100 000 clip subset of the *Yahoo! Flickr Creative Commons 100 Million (YFCC100M) dataset to supplement the test set
- MED 2017
 - Phased out the Heterogeneous Audio Visual Internet Collection (HAVIC) Progress portion of the test set; HAVIC development resources still provided to teams
 - Added an additional 100 000 clips from YFCC100M to the test set
 - Using last years Ad-Hoc events as this years Pre-Specified events
 - Added 10 new Ad-Hoc events; with exemplars from the YFCC100M dataset
 - Dropped support for several evaluation conditions

* - Disclaimer: Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

MED '17 Events

Pre-Specified Events	Ad-Hoc Events
MED '16 AH Events	New Events
Camping	Fencing
Crossing a Barrier	Reading a Book
Opening a Package	Graduation Ceremony
Making a Sand Sculpture	Dancing to Music
Missing a Shot on a Net	Bowling
Operating a Remote Controlled Vehicle	Scuba Diving
Playing a Board Game	People Use a Trapeze
Making a Snow Sculpture	People Performing Plane Tricks
Making a Beverage	Using a Computer
Cheerleading	Attempting the Clean and Jerk

Fencing

Definition:

Two individuals fight with swords according to a set of rules

Explication:

Fencing is the Olympic sport of sword fighting. Fencing consists of swings, dodges, or parries, in order, to either avoid getting hit by the opponent's sword or in an attempt to strike the opponent with your sword. People not using the proper equipment (wire guard mask and sword) are not considered fencing. Only matches between two individuals are considered positive for this event, though multiple simultaneous one-on-one matches can co-occur...

Evidential Description:

- scene: outside or inside, but usually in a gym
- objects/people: foils, épées, or sabers; protective fencing gear, such as wire guard mask and padded suits; sometimes boundary lines on floors
- activities: standing, swinging/thrusting swords, dodging, and parrying
- audio: sounds of swords hitting swords or bodies; crowd cheering

Illustrative Examples

- Positive instances of the event
- Non-Positive “miss” clips that do not contain the event



Miss →

Ad-Hoc Event Creation

- Ad-Hoc exemplars from YFCC100M, which is unannotated
 - First time sourcing exemplars from YFCC100M
- Using an Aladdin system from 2016 we performed the following
 1. Selected exemplars for candidate events from HAVIC
 2. Trained the system, then searched YFCC100M
 3. Selected exemplars from the top 200~400 results, prioritizing diversity

Test Data

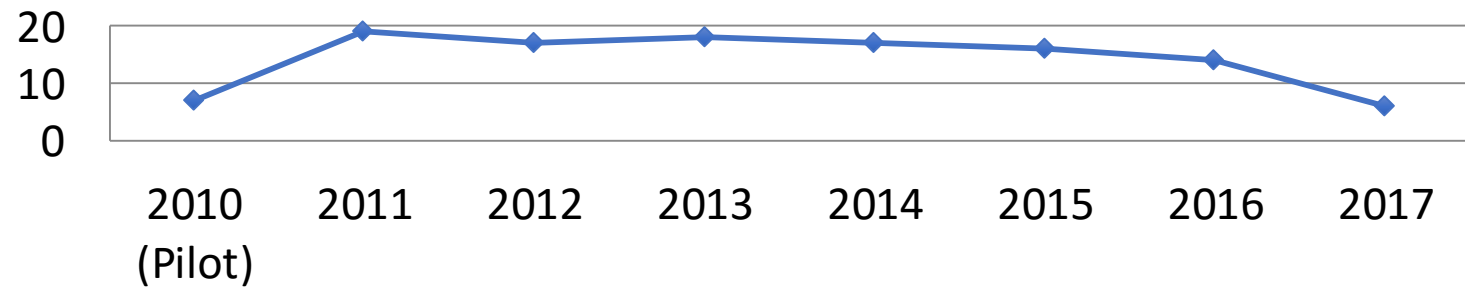
- MED '17 discontinued the use of the HAVIC Progress set for evaluation
- Additional YFCC100M Subset
 - Random selection* (Same criteria as the MED16 YFCC100M subset)
- MED '17 required processing the full 2 050 hour dataset (referred to as MED17EvalFull)
 - Full dataset for MED '16 (MED16EvalFull) was 4 738 hours

Data collection	# of videos	Duration (h)	Avg. duration (s)
MED16 YFCC100M Subset	100 000	1 025	37
MED17 YFCC100M Subset	100 000	1 025	37
Total (MED17EvalFull)	200 000	2 050	37

* - Excluding YLI-MED corpus videos (~50k videos) [Bernd et al. The YLI-MED Corpus: Characteristics, Procedures, and Plans; ICSI Technical Report TR-15-001]; Excluding videos not available by mmcommons.org's Amazon Web Services (AWS) Simple Storage Service (S3) data store (~5k videos)

6 MED 2017 Finishers By Condition

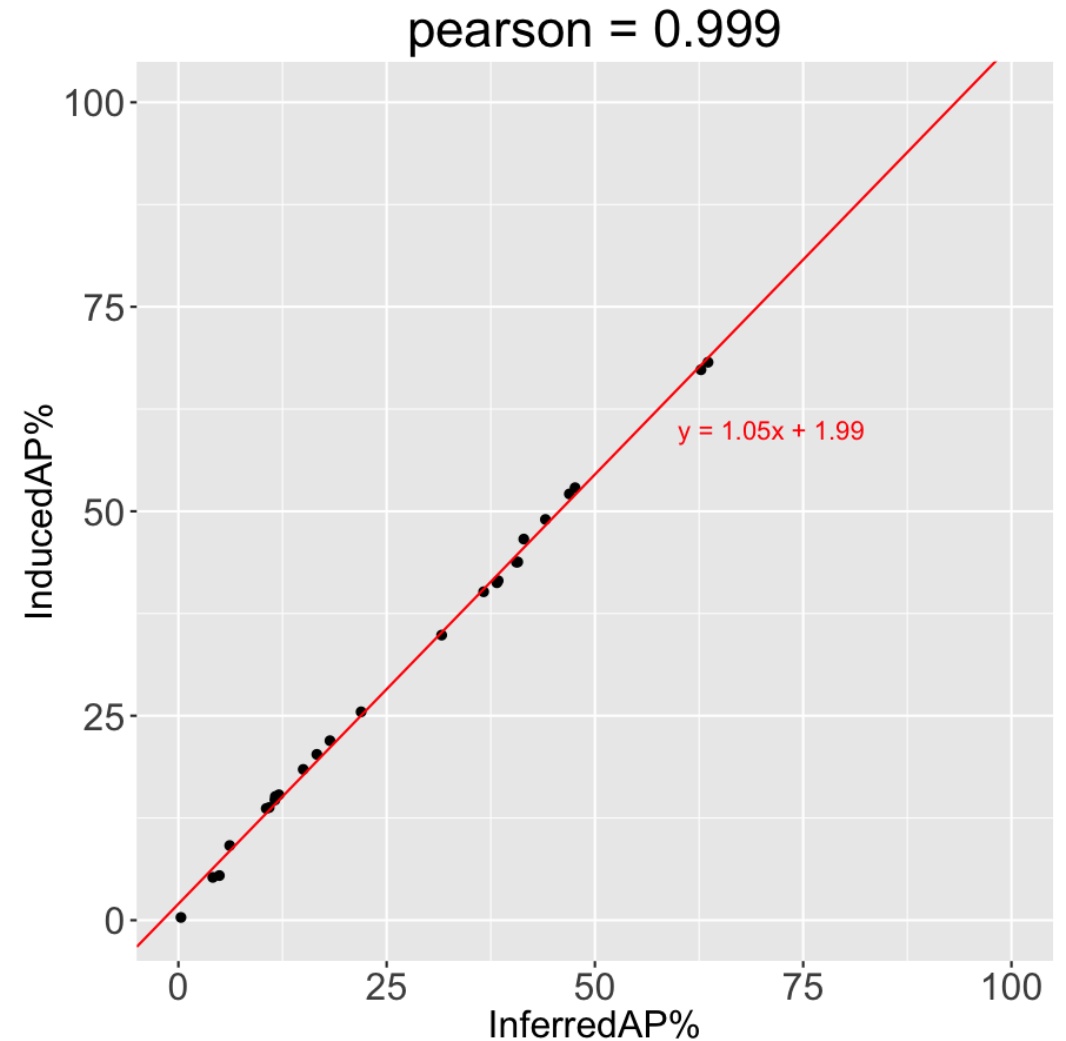
Years	Team	Pre-Specified	Ad-Hoc	Organization
7	INF	✓	✓	Carnegie Mellon University et al.
	MediaMill	✓	✓	MediaMill - University of Amsterdam
	TokyoTech	✓	✓	Tokyo Institute of Technology, National Institute of Advanced Industrial Science and Technology
4	ITICERTH	✓	✓	Informatics and Telematics Inst.
	MCISLAB	✓		Beijing Institute of Technology Mcislab
3	BUPTMCPRL	✓		Multimedia Communication and Pattern Recognition Labs, Beijing University of Posts and Telecommunications
		6	4	



◆ Number of MED Finishers

Metric – Inferred Average Precision

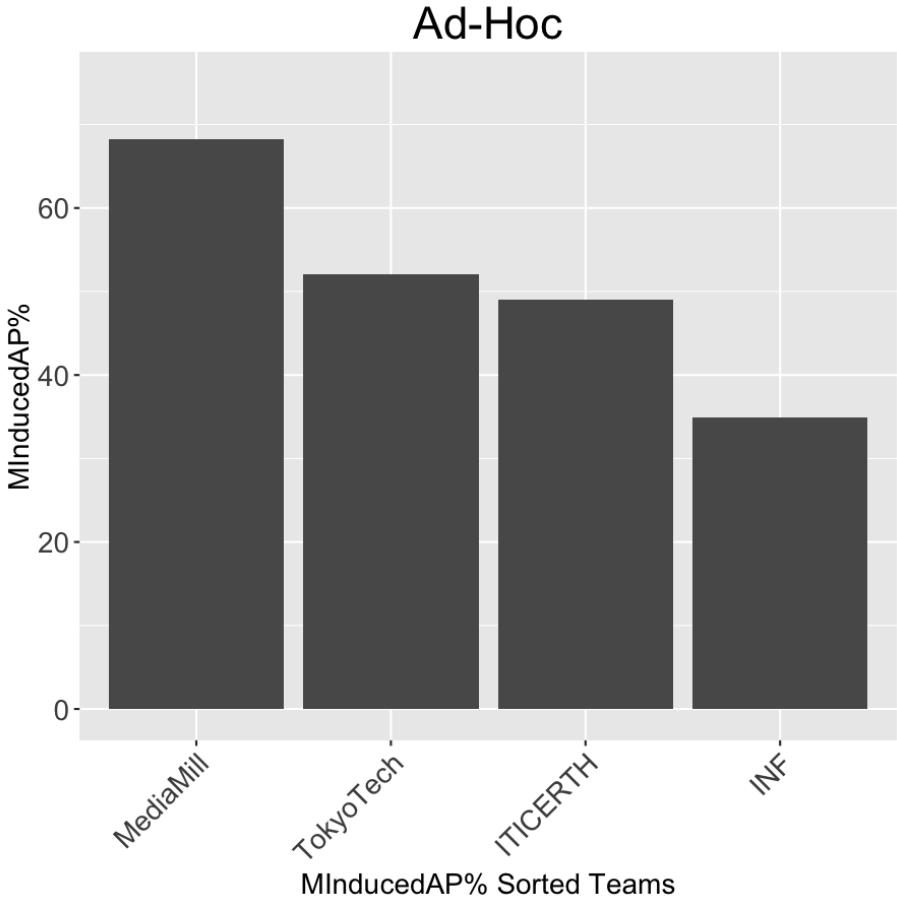
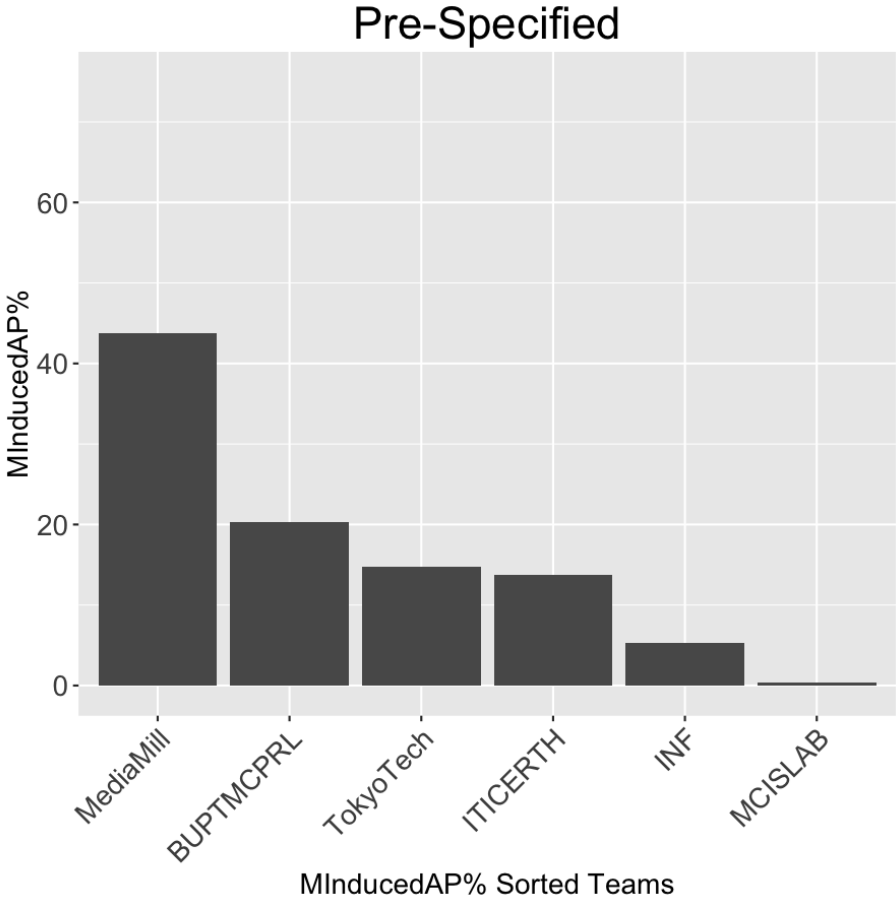
- Inferred Average Precision - Follows Aslam et al.^[1] procedure to approximate Average Precision using stratified, variable density, pooled assessment
- For MED '15, NIST ran experiments with 2014 data to optimize the strata sizes and sampling rate. This same sampling rate was used for MED '16, and again for MED '17
 - Define 2 strata
 - 1-60 -> 100 %
 - 61-200 -> 20 %
- **Due to a misconfiguration of our scoring pipeline, we've actually been reporting Induced Average Precision (InducedAP)**



[1] - Aslam et al. Statistical Method for System Evaluation Using Incomplete Judgments; Proceedings of the 29th ACM SIGIR Conference, Seattle, 2006.

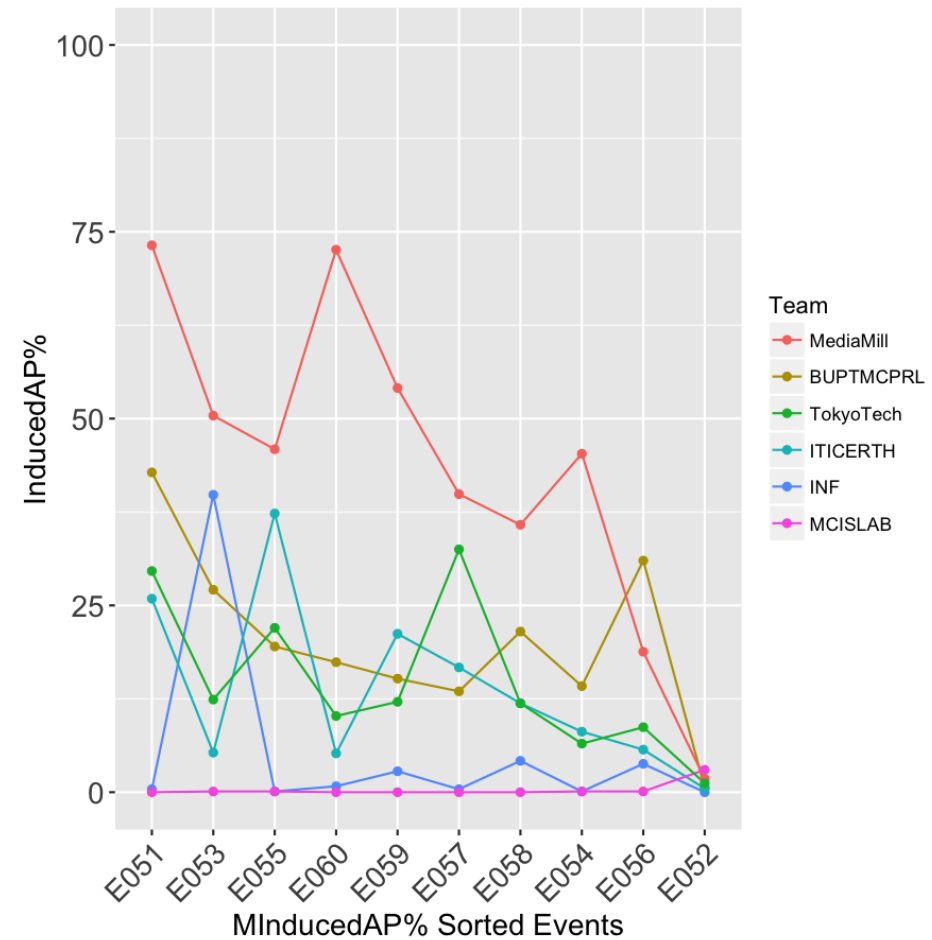
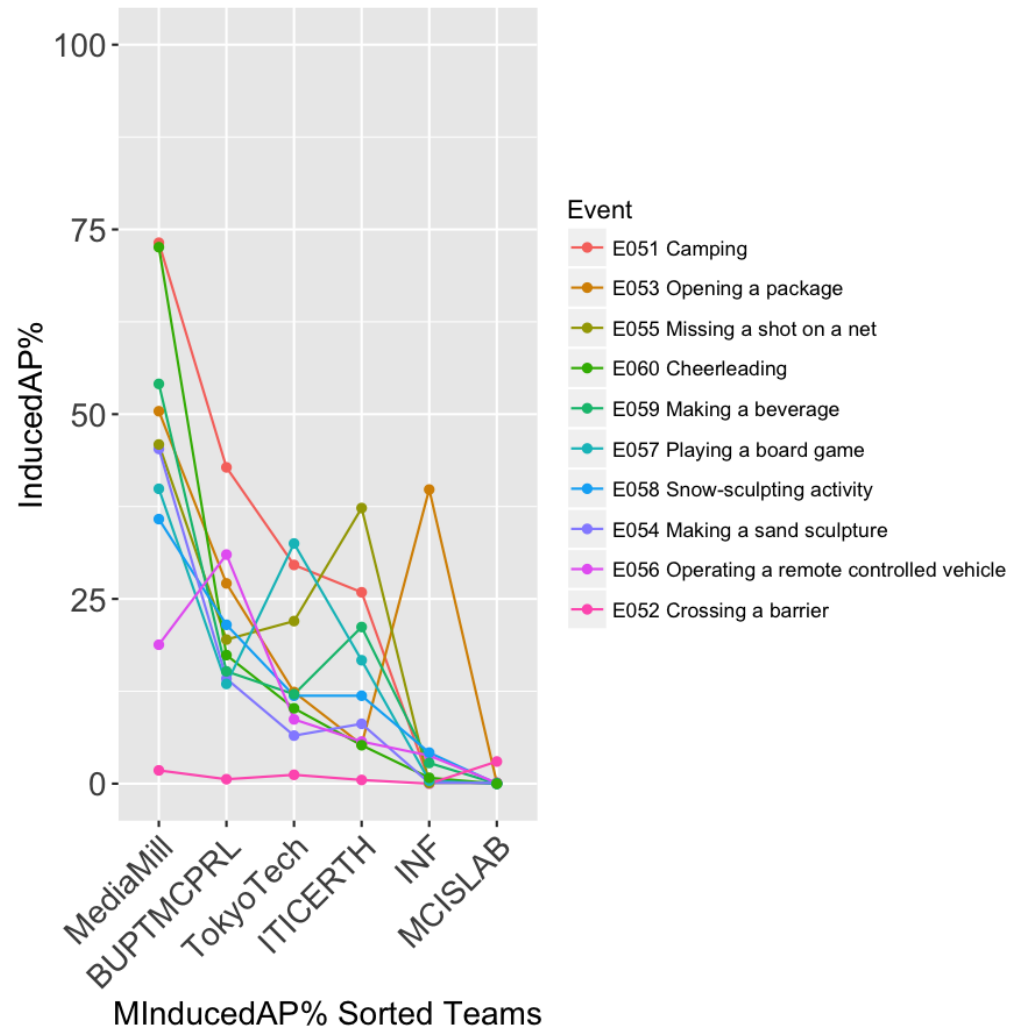
Mean InducedAP (MInducedAP) Across Events

Results of Primary Systems



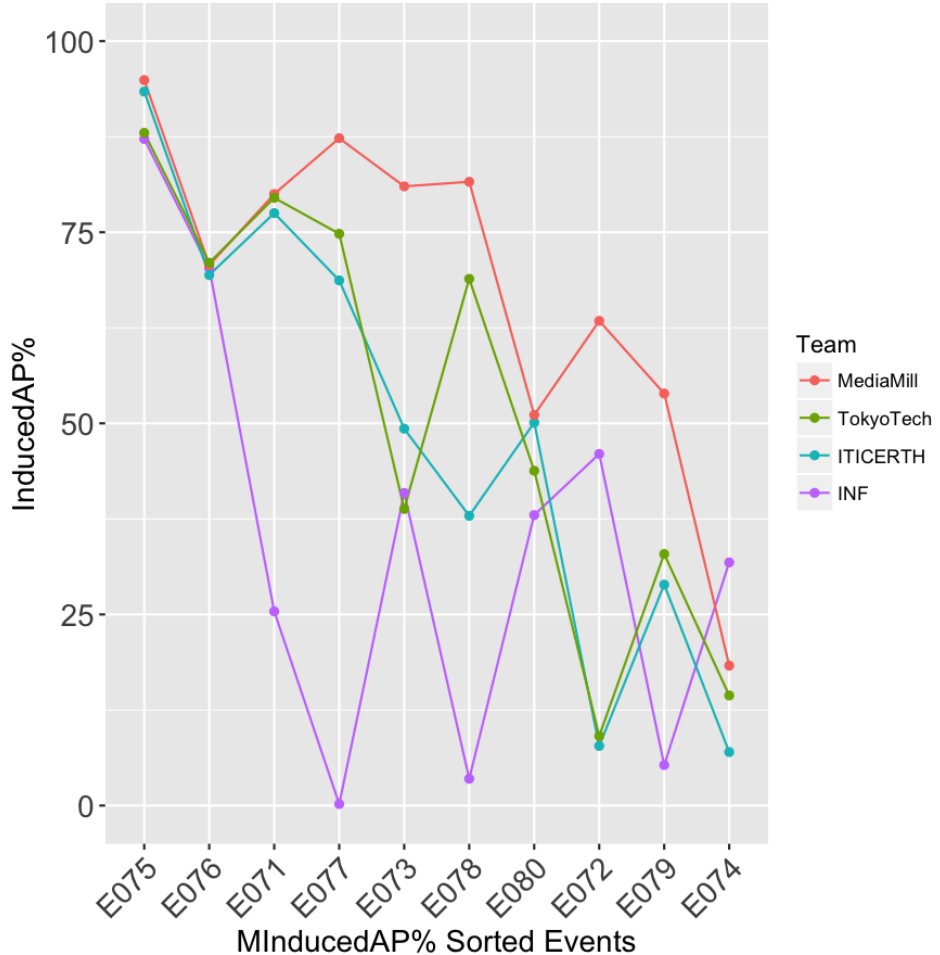
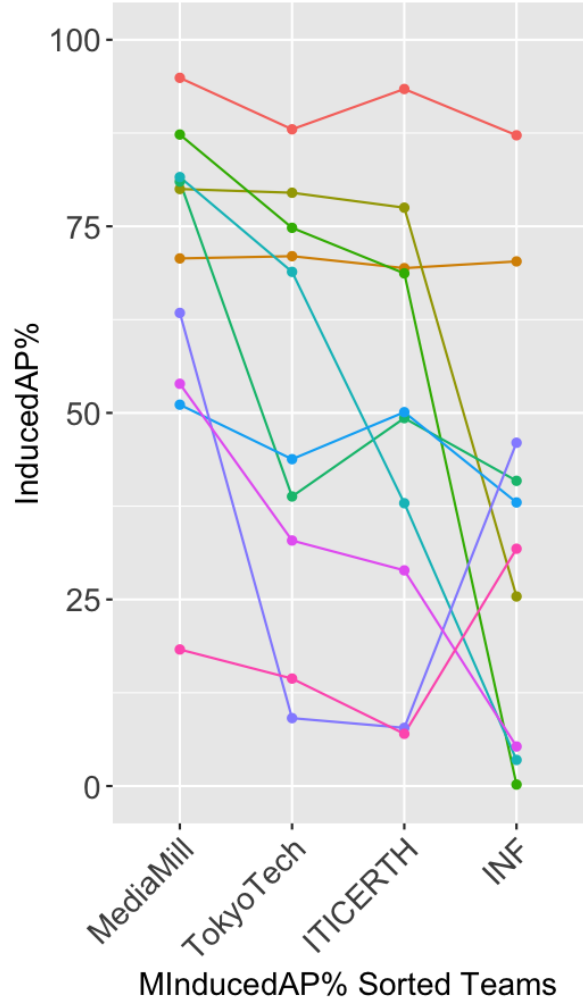
Pre-Specified InducedAP by System and Event

Primary Systems



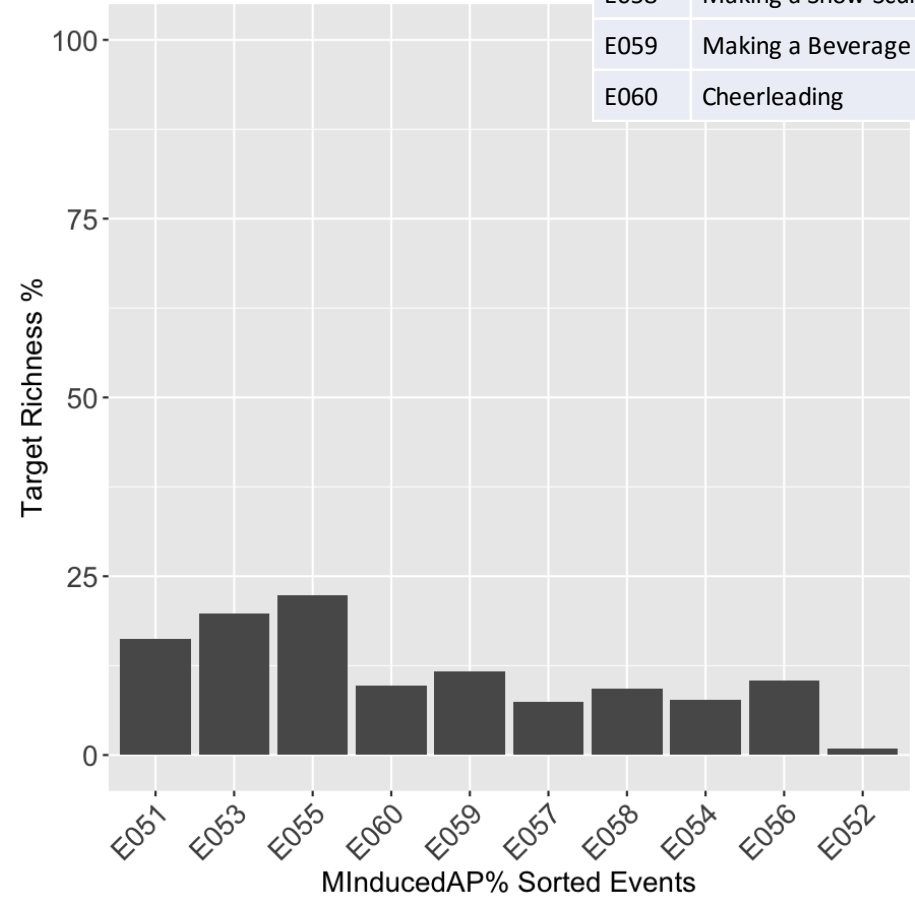
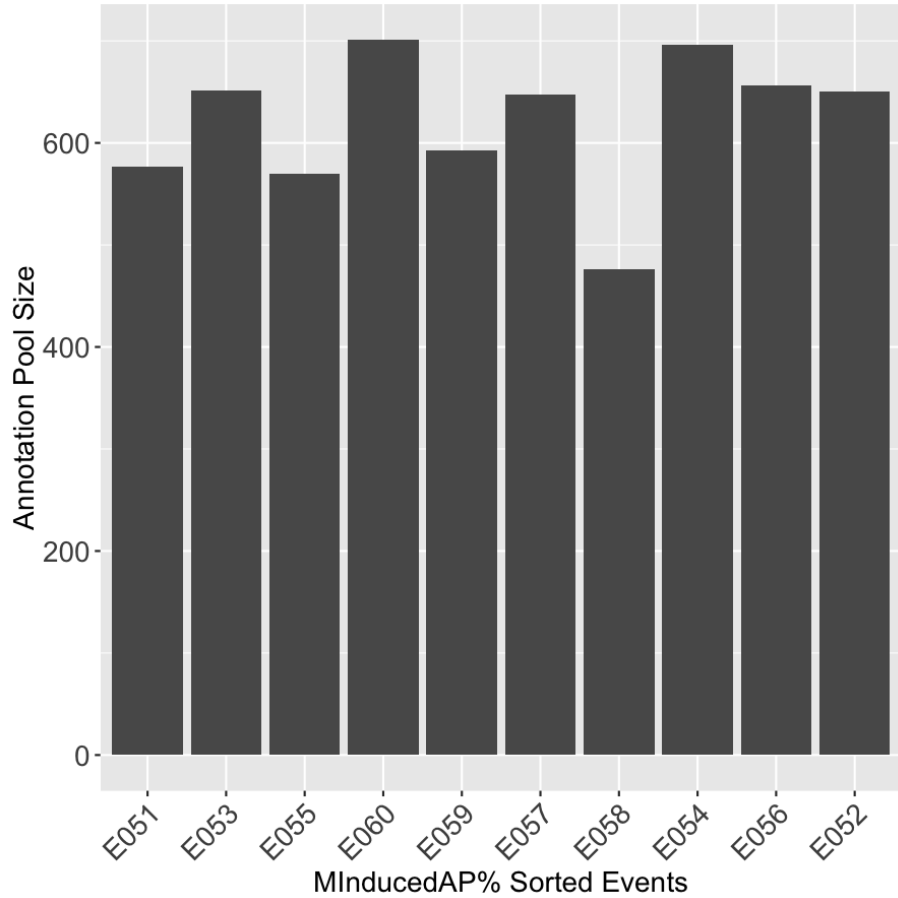
Ad-Hoc InducedAP by System and Event

Primary Systems



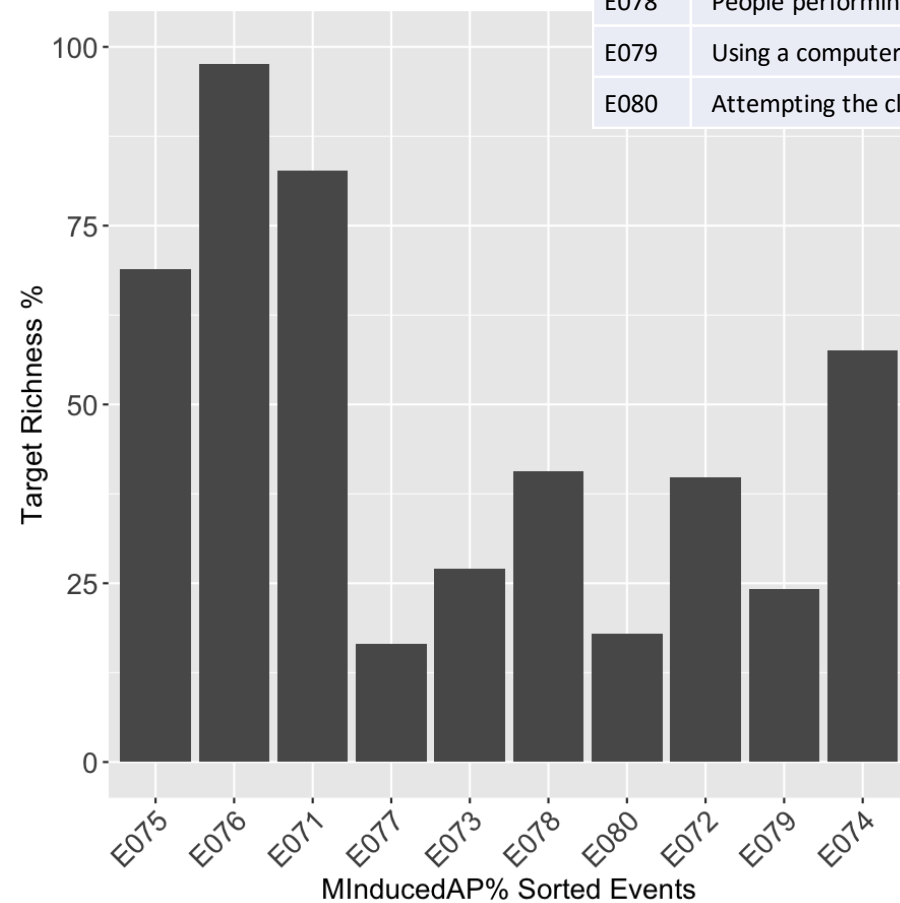
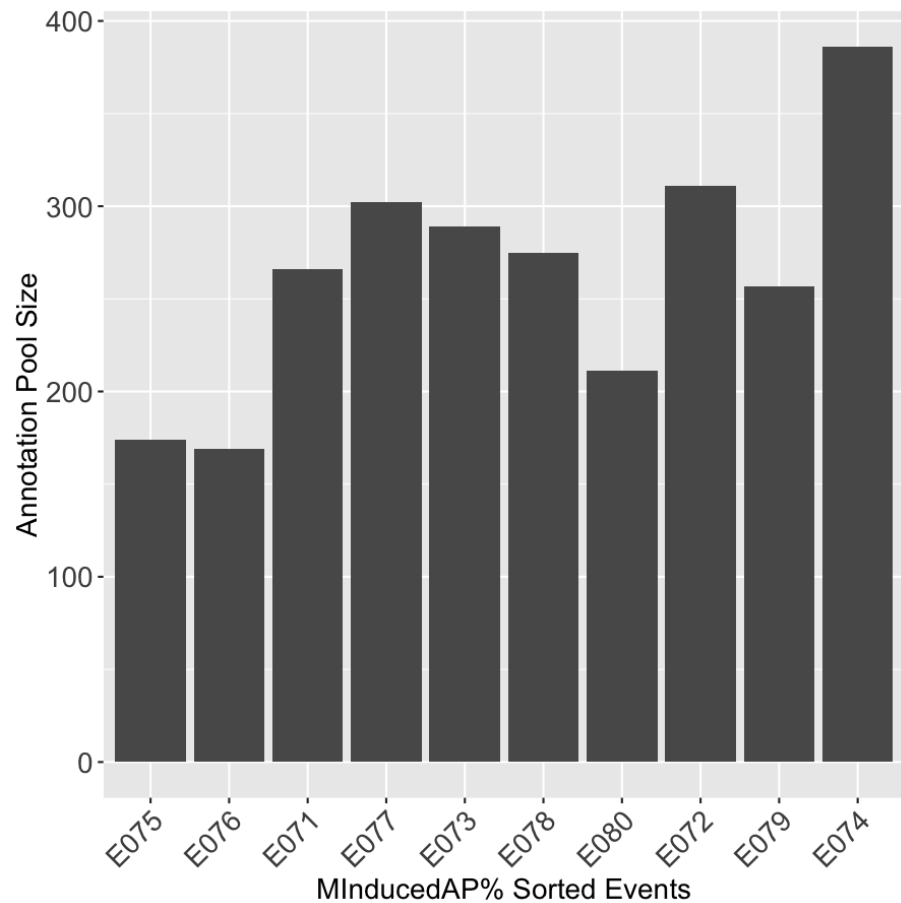
E051	Camping
E052	Crossing a Barrier
E053	Opening a Package
E054	Making a Sand Sculpture
E055	Missing a Shot on a Net
E056	Operating a Remote Controlled Vehicle
E057	Playing a Board Game
E058	Making a Snow Sculpture
E059	Making a Beverage
E060	Cheerleading

Pre-Specified Pool Size and Target Richness



Ad-Hoc Pool Size and Target Richness

E071	Fencing
E072	Reading a book
E073	Graduation ceremony
E074	Dancing to music
E075	Bowling
E076	Scuba diving
E077	People use a trapeze
E078	People performing plane tricks
E079	Using a computer
E080	Attempting the clean and jerk



MED '17 Summary

- Noticeable drop in participation this year
- All teams built a “Small” hardware system
- Different datasets and exemplar selection process
- Target richness for some AH events approaching 100%

MED '18 Plans

- Progress annotations to be released shortly after TRECVID 2017
- If we continue MED for 2018, what might it look like?
 - Bring back support for a “Sub” test set (e.g. MED16EvalSub)?
 - Bring back the 0 Exemplar evaluation condition?
 - Subdivide SML hardware condition?
 - Update the Ad-Hoc exemplar scouting procedure?
- Thoughts?

Thank you!

Questions?